Clustering: Agglomerative Hierarchical Clustering

by xz431

Problem Submissions Leaderboard Discussions

In this programming assignment, we're going to implement the agglomerative hierarchical clustering algorithm. We will implement three different cluster similarity measures: single link, complete link, and average link.

We will be using geographical data for this assignment. Each of the data points is a 2-D vector, with longitude and latitude as its dimensions.

Here is a brief review of the agglomerative clustering. During the clustering process, we iteratively aggregate the most similar two clusters, until there are K clusters left. For initialization, each data point forms its own cluster.

The similarity of two clusters C_i, C_j is determined by a distance measure. We use the following three measures in this assignment:

Single link:

$$D(C_i, C_j) = \min\{d(\mathbf{v}_p, \mathbf{v}_q) \mid \mathbf{v}_p \in C_i, \mathbf{v}_q \in C_j\}$$

Complete link:

$$D(C_i, C_j) = \max\{d(\mathbf{v}_p, \mathbf{v}_q) \mid \mathbf{v}_p \in C_i, \mathbf{v}_q \in C_j\}$$

Average link:

$$D(C_i, C_j) = \text{mean}\{d(\mathbf{v}_p, \mathbf{v}_q) \mid \mathbf{v}_p \in C_i, \mathbf{v}_q \in C_j\}$$

The smaller the distance is, the more similar the two clusters are.

In the equations, $d(\cdot, \cdot)$ is a distance measure between two data points. In this assignment, for simplicity, we use the Euclidean distance, defined by:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_i (p_i - q_i)^2}$$

where p_i, q_i are dimensions of $\mathbf{p}, \mathbf{q}$.

You're encouraged to think of a better distance measure than Euclidean distance for our data.

There's one data point testing your algorithm's efficiency on *single link* clustering. Please think about the time complexity of your algorithm.

Input Format

The first line of the input will be three space separated integers N K M:

1. The number of data points (lines) following the first line N.

2. The number of output clusters K.

3. The cluster similarity measure M to be used. $M = 0$ for single link, $M = 1$ for complete link, $M = 2$ for average link.

Starting from the second line, each line will have exactly two floating point numbers, representing longitude and latitude of a location. Each line corresponds to a 2-D data point which will be fed into the clustering algorithm.

There will be $N + 1$ lines in total.

Our objective is to use the similarity measure specified by M to cluster N data points into K clusters.

Constraints

Machine learning libraries are not allowed in this assignment. For a complete list of allowed libraries, please refer to the column *standard challenges* on this page.

Output Format

There should be N lines in the output, and one integer for each line, indicating the cluster label of the data point. The i-th line in the output corresponds to the i-th data point in the input, i.e. the $(i + 1)$-th line in the input.

The exact numbers used for cluster labels don't matter, but there should be exactly K unique labels, and the data points in the same cluster should share the same label.

Sample Input 0

```
5 2 0
51.5217 30.1140
27.9698 27.0568
10.6233 52.4207
122.1483 6.9586
146.4236 -41.3457
```

Sample Output 0

```
0
0
0
1
1
```

Explanation 0

The first line of the input shows that there will be 5 data points, the objective is to cluster them into 2 clusters, and the cluster distance metric to be used is single link.

The output shows that the first three data points should be of the same cluster, while the other two points belongs to the other cluster.

Submissions: 174
Max Score: 5
Difficulty: Medium

Rate This Challenge:

More

Current Buffer (saved locally, editable) Python 3

1

Upload Code as File

Run Code

Submit Code